유전자 발현 데이터를 활용한 인공지능 기반 유방암 아형 예측 및 바이오마커 탐색

한지연[1], 박세연[2], 김현희[3]
[1]동덕여자대학교 정보통계학과 학부생
[2]동덕여자대학교 응용화학과 교수
[3]동덕여자대학교 정보통계학과 교수

choojh7@naver.com, sypark21@dongduk.ac.kr, heekim@dongduk.ac.kr

AI-based Prediction of Breast Cancer Subtypes and Biomarker Discovery Using Gene Expression Data

Ji-Yeon Han[1], Seyeon Park[2], Hyon Hee Kim[1]
[1]Dept. of Statistics and Information Science, Dong-Duk Women's University
[2]Dept. of Applied Chemistry, Dong-Duk Women's University

요 약

본 연구는 유전자 발현 데이터를 활용하여 유방암 환자의 분자 아형을 보다 정밀하게 분류하고, 새로운 바이오마커를 탐색할 수 있는 인공지능 기반 아형 분류 모델을 제안한다. 이를 위해 TCGA-BRCA RNA-seq 데이터를 활용하고, PAM50 에서 정의된 아형을 라벨로 설정하여 XGBoost 기반 다중분류 모델을 구축하였다. 모델은 학습과 검증에서 91.1% 정확도를 달성하여 기존 면역조직화학적 분류법의 한계를 보완할 수 있음을 확인하였다. 또한 특징 중요도 분석을 통해 PAM50 과 일치하는 핵심 유전자를 재현함과 동시에 FSCN1, TPX2, CDCA8 등 새로운 바이오마커 후보를 제시하였다. 이러한 결과는 인공지능 기반 접근이 유방암의 정밀한 분자 아형 분류와 맞춤형 치료 전략 수립에 기여할 수 있음을 보여준다.

1. 서론

유방암은 전 세계 여성에게 가장 흔히 발생하는 암 중 하나이며, 치료 방법과 환자의 예후 평가는 아형 분류에 크게 의존한다. 현재 임상에서는 에스트로겐 수용체(ER), 프로게스테론 수용체(PR), HER2 발현 여부 등 면역조직화학법(IHC)에 기반한 아형 분류가 널리 활용되고 있으나, 이는 분자적 특성을 직접 반영하지 못한다는 한계가 있다.

이러한 한계를 극복하기 위해 RNA 발현 데이터를 활용한 유전자 발현 데이터를 활용한 분자 아형 분류법이 제안되었으며, 그 대표적인 예가 PAM50 이다. PAM50 은 50 개의 핵심 유전자를 활용하여 다섯 가지 아형을 정의하며, IHC 보다 예후 예측력이 우수하다고 보고된 바 있다[1]. 그러나 PAM50 은 제한된 유전자 집합과 단순 규칙 기반 접근에 의존하기 때문에, 대규모 발현 데이터에 내재된 복잡한 분자적 상호작용과 비선형 패턴을 충분히 반영하지 못하는 한계가 있다.

이에 본 연구에서는 약 2 만여 개 유전자를 모두 고려한 인공지능 기반 아형 분류 모델을 구축하고, PAM50 과 비교하여 그 보완 가능성을 탐색하고자 한다. 또한 특징 중요도 분석을 통해 PAM50 과 일치하는 유전자를 확인하였고 동시에 새로운 바이오마커 후보를 제시하였다. 이러한 결과는 인공지능 기반 유방암 아형 분류 모델이 향후 임상 적용 및 분자 의학 연구의 확장 가능성을 제시한다.

2. 데이터 수집 및 전처리

본 연구에서는 UCSC Xena 플랫폼에서 제공하는 TCGA-BRCA (Breast Invasive Carcinoma) 데이터셋을 활용하였다. RNA-seq 기반 유전자 발현 데이터를 사용하였으며, 해당 데이터는 log2(x+1) 변환 및 정규화가 이미 적용되어 있어 추가적인 스케일링 과정은 수행하지 않았다.

임상 데이터(Clinical data)에는 PAM50 알고리즘 기반 아형 분류 정보가 포함되어 있으며, 이를 본 연구의 타깃 변수로 사용하였다. PAM50 아형은 다섯 가지로 구분되지만, Normal-like 아형은 샘플 수가 적

고 분류 신뢰성이 낮으며 임상적 활용도도 제한적이어서 제외하였다[2]. 따라서 최종적으로 네 가지 아형(HER2, Luminal A, Luminal B, Triple Negative)을 분석 대상으로 정의하였다. 또한, 용어의 일관성을 위해 "HER2-enriched"는 "HER2", "Basal-like"는 "Triple Negative"로 명칭을 통일하였다.

데이터 정제 과정에서 아형 정보가 결측된 환자 샘플은 제거하였으며, 최종적으로 837 명 환자 샘플이 확보되었다. 각 환자에 대해 20,530 개 유전자 발현 값이 독립 변수로 포함되었고, 아형 라벨을 종속 변수로 사용하였다. 이후 모델 학습을 위해 전체 데이터를 학습용과 평가용으로 분할하였으며, 라벨 분포를 고려하여 stratified train/test split 을 적용하여 학습 80%, 평가 20%로 나누었다. 아형 간 일부 불균형이 존재하였으나, 불균형 처리 시 성능 저하가 관찰되어 원본 분포를 유지한 상태로 학습을 진행하였다. 아형별 분포는 <표 1>에 제시하였다.

	HER2	Luminal A	Luminal B	Triple Negative
Counts	434	194	142	67

<표 1> TCGA-BRCA 데이터셋의 유방암 아형 분포

3. 분자 아형 예측 모델 개발

본 연구에서는 유방암 분자 아형을 예측하기 위해 고차원 · 비선형적 특성을 지닌 RNA-seq 발현 데이터를 효과적으로 학습할 수 있는 그래디언트 부스팅 기반의 XGBoost 다중분류 모델을 구축하였다. 입력 변수는 RNA-seq 발현 데이터이며, 출력 라벨은 임상 데이터에서 제공된 Luminal A, Luminal B, HER2, Triple Negative 네 가지 아형이다.

모델 학습 과정에서 하이퍼파라미터 최적화를 위해 RandomizedSearchCV 를 적용하였으며, 최적화 결과 n_estimators=300, learning_rate=0.01, max_depth=8, subsample=0.619, colsample_bytree=0.414, gamma=0 의 조합이 최적 파라미터로 선정되었다.

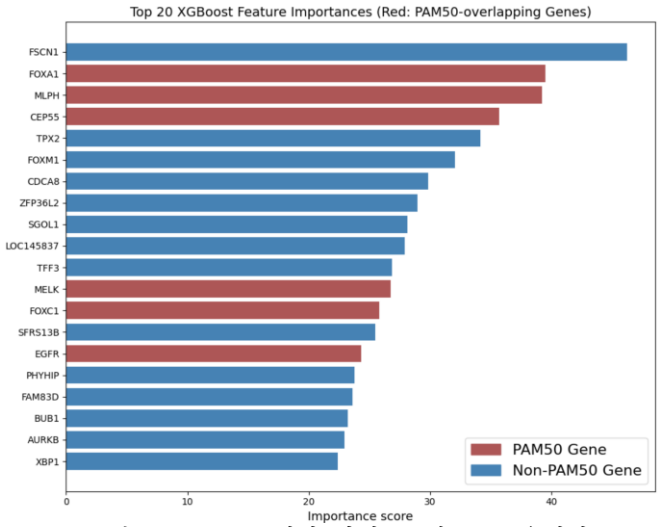
평가 데이터셋(20%)으로 성능을 검증한 결과, 전체 정확도는 91.1%로 나타났으며, 평균 F1-score 는 0.89 로 나타났다. 아형별 성능지표는 <표 2>에 제시하였다.

	Precision	Recall	F1-score
HER2	0.85	0.85	0.85
Luminal A	0.89	0.99	0.93
Luminal B	0.94	0.74	0.83
Triple Negative	1.00	0.93	0.96

<표 2> Classification Report

4. 설명가능 인공지능 기술 적용

본 연구에서는 학습된 XGBoost 모델의 해석 가능성을 확보하기 위해 feature importance 분석을 수행하였다. <그림 1>은 상위 20 개의 중요 유전자를 나타내며, PAM50 패널에 포함된 유전자는 빨간색으로 표시하였다. 추가적으로 중요도 상위 50 개 유전자를 기준으로 PAM50 과의 교집합을 확인한 결과, FOXA1, MLPH, CEP55 등 총 10 개의 유전자가 일치하는 것으로 나타났다. 이는 제안된 모델이 기존 PAM50 분류법에서 강조된 핵심 분자적 특징을 재현했음을 시사한다. 동시에 FSCN1, TPX2, CDCA8 등 PAM50 에 포함되지 않았으나 높은 중요도로 도출된 유전자들도 확인되었으며, 이는 향후 새로운 바이오마커 후보로서의 가능성을 보여준다.



<그림 1> XGBoost 기반 상위 20 개 중요 유전자

5. 결론

본 연구에서는 TCGA-BRCA RNA-seq 발현 데이터를 활용하여 유방암 분자 아형을 분류하는 XGBoost 기반 인공지능 모델을 구축하였다. 제안된 모델은 다중 아형 분류에서 높은 수준의 예측 성능을 보였으며, feature importance 분석을 통해 PAM50 과 중복되는 핵심 유전자를 확인함으로써 기존 분류 체계의 타당성을 뒷받침하는 동시에 새로운 후보 유전자를 제안하였다.

이러한 결과는 인공지능 기반 접근이 기존 분류법의 한계를 보완하고, 유방암 환자의 분자 진단 및 맞춤형 치료 전략 수립에 기여할 수 있음을 시사한다. 또한 본 연구는 새로운 바이오마커 탐색의 가능성을 열어주며, 분자 의학적 연구의 확장으로 이어질 수 있다. 향후 연구에서는 다기관의 데이터셋을 활용하여 모델의 일반화 가능성과 임상 적용성을 더욱 강화할 예정이다.

참고문헌

[1] T. O. Nielsen et al., "A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor positive breast cancer," Clinical Cancer Research, vol. 16, no. 21, pp. 5222-5232, 2010.Clinical Cancer Research, vol. 16, no. 21, pp. 5222-5232, 2010.

[2] A. Prat, J. S. Parker, O. Karginova, C. Fan, C. Livasy, J. I. Herschkowitz, X. He, and C. M. Perou, "Deconstructing the molecular portraits of breast cancer," Molecular Oncology, vol. 5, no. 1, pp. 5-23, 2011.